ED 291 805 MISINTERPRETING AND MISUSING TESTS: SOME NEW WAYS.
GREEN, DONALD ROSS
1 OF 1 5 MAY 85 5P.

DOCUMENT RESUME

ED 291 805 TM 011 123

AUTHOR Green, Donald Ross
TITLE Misinterpreting and Misusing Tests: Some New Ways.
PUB DATE 5 May 85
NOTE 5p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Gains; *Item Banks; *Latent Trait Theory; *Standardized Tests; *Test Construction; Testing Programs; *Test Interpretation; Test Items; Test Norms; *Test Use

ABSTRACT

The use of item banks and item response theory has resulted in new ways to misinterpret and misuse tests through customized, yet standardized, achievement test batteries. The new test batteries create the possibility of serious misunderstandings based on the idea that any subset of items from the pool with a proper range of difficulties will produce scores that can reasonably be interpreted using national norms. System personnel may believe that they can select items from the nationally normed pool in accord with their preferences. Items chosen to be "fair" to a local area will provide higher scores than will a representative sample. The effect of local scales cannot be ignored in comparing local and national norms. If carefully used, the new tests can provide useful information. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DONALD ROSS GREEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Misinterpreting and Misusing Tests: Some New Ways

by

Donald Ross Green
CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED291805

Traditionally problems of test use have arisen when ill informed people have overinterpreted test scores, or misunderstood them, or used them for purposes not intended by the authors and publishers. Recently some new versions of these old problems have surfaced stemming from the use of item banks and item response theory.

Together these two more or less new technologies have been used to produce the new wonder of the age: The Customized (yet) Standardized Achievement Battery.

This new marvel is constructed so that it:

1) matches the local curricula precisely.
2) provides mastery information on each objective in those curricula.
3) provides data for assessment of status and growth via national norms.
4) is appropriate for Chapter 1 evaluation, for selecting students for gifted programs, and for planning IEPs.
5) can be equated to the tests previously used, whatever they may have been.
6) can be modified at will without the need to do further equating studies.

This is a minimal list of the criteria for these new wonders, as the examples of RFPs offered by Hoover(1985) indicate, but it may be enough to show what problems can arise. There are, of course, good reasons for asking for such features. They arise from the wish to:

1) increase curricular relevance and satisfy teachers' complaints that standardized tests do not measure what they are teaching.
2) provide teachers and curriculum supervisors with criterion referenced achievement data.
3) provide all the information that administrators, school boards and the public have come to expect from traditional standardized test batteries.
4) eliminate multiple testing programs.
5) maintain continuity of data with the past.
6) be able to modify the tests readily (after all, curricula emerge) without losing any of the preceding benefits.

Each of these taken singly is a reasonable goal although the last one has not seemed possible heretofore. Now however not only is the last one judged possible but also it is widely believed that all the others are

TM 011 123

possible simultaneously from just one test. The new problems arise from the fact that under certain circumstances that belief is almost correct. Many if not most people involved in these undertakings seem to be a bit vague about the difference between almost correct and correct as well as unaware that the "certain circumstances" may not exist.

Given a large item bank in which the items have been calibrated on nationally normed scales using item response theory many things are indeed possible. A large and varied array of such items in the item bank and what Wright (1977) has called the "sample-free item calibration" and "item-free person measurement" feature of IRT procedures do make a powerful technology. They do facilitate the development of tests that can provide both criterion referenced and normative data, they do make it possible to insert, delete, or substitute items while maintaining the integrity of the scale, they do sometimes facilitate equating, and they even may make it possible to build a test that will serve several purposes traditionally met with separate tests.

However none of this is automatic and the attempt to do all these things at once can create confusion and misunderstanding which are the building blocks of misinterpretation and misuse. Especially because these technologies are new to many people a number of the possible misinterpretations and misuses become likely. Perhaps most noteworthy are those that arise from siren song of item-free measurement which can lead to several new sorts of misunderstandings. A hypothetical example may help make the dangers clear.

Suppose a school system, about to inaugurate a new curriculum, contracts for one of the new wonder test batteries which is to have the features listed above. Since naturally their new curriculum has many unique features, the battery incorporates new items written to the local system specifications along with a set of items already calibrated to a nationally normed scale. It is planned to calibrate these new items on the normed scale and to repeat this process each time the test is given thus building a large pool of calibrated items which focus on the objectives found in the local curriculum.

This situation creates the possibility of at least two serious misunderstandings. Both are based upon the notion that measurement of student performance is truly item free, i.e., that any subset of items from this pool with a proper range of locations (difficulties) will produce scores that can reasonably be interpreted using the national norms.

First, this idea is likely to lead the system personnel to believe that they can select the items from the nationally normed pool in accordance with their preferences. This is one road to instant "improvement" in test scores. This will happen because it appears to be an article of faith in many educational circles, including teachers and curriculum personnel, that the item is unfair if doesn't ask what was directly taught and in the format it was taught. Thus only "fair" items tend to be picked by local personnel even though these conditions did not apply to the national standardization sample. Consequently items chosen to fit those criteria will produce higher scores in the local sample than would a sample of items generally representative of the trait that has been normed. Of course the publisher should resist this pressure to distort the tests.

Let us assume that the resistance was successful; CTB has found it possible to do this without losing the contract albeit with difficulty. Now also assume that a year passes during which the teachers vigorously present the new curriculum and the test, with or without further new items, is given again. All concerned, teachers and administrators alike, now look forward eagerly to see how successful they have been in implementing this new curriculum.

What will they find? A sharp improvement on the normed scales seems almost inevitable. Does this mean that the instruction was successful? Of course it does; focussed instruction works as everyone knows. Does this also mean that the system has improved relative to others. Of course it does not, "improved" percentile ranks notwithstanding. A stacked deck is a stacked deck even if it is called a customized test.

The measure of the deception, [illegible]core misinterpretation if you prefer, can be seen in the difference betwe[illegible]he norms estimates obtained from the nationally calibrated items and t[illegible]se obtained from the items calibrated locally using the data from the first testing. Given enough items of each sort (say twenty or more), two norms estimates can be obtained. If the local curriculum is indeed unique (if it is not, the taxpayers' money has been ill spent) and if the teachers indeed taught it (also not a sure thing), the items developed specifically to match the local curriculum should show greater gains than those that fairly represent the nationally normed trait.

Thus there are now two scales, a local scale and a national scale. To conceal this by not examining the data for such effects is clearly a misuse of the test. On the other hand to report both should be a useful way to proceed. Obviously when this is done great care will have to be taken to avoid confusion about the meaning of the two numbers, especially since they will look alike unless one of them is transformed. However if the local curriculum has any unique value these data should provide useful information about it to all concerned.

This example has been used to show that the use of item response theory does not make achievement tests curriculum-and/or instruction-free; the term item-free measurement refers to procedures that can be used only in certain limited circumstances. Similarly the results of local calibrations cannot be used elsewhere, i.e., item calibrations are sample free also in a limited sense only.

Clearly the examples just discussed do not exhaust the list of possible new misinterpretations and misuses that may arise and of course all the old ones remain. Add the pitfalls of context effects and the potential problems created by the lack of unidimensionality described by others in this symposium and it may be seen that the possibilities of test misinterpretation and misuse have indeed been expanded by these technological innovations.

Nevertheless it may also be seen that, like all tests, when carefully used the new wonder tests can provide useful information not otherwise obtainable.

References

Hoover, H.D., Are there limits to test use? Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1985.

Wright, B.D., Solving measurement problems with the Rasch Model, Journal of Educational Measurement, 1977, 14, 97-116.